IN THE SPECIFICATION

Please replace the Title as follows:

~~*A Program for Microarray Design and Analysis*~~

A1 *A Computer-Based Method for Creating Collections of Sequences from a Dataset of Sequence Identifiers Corresponding to Natural Complex Biopolymer Sequences and Linked to Corresponding Annotations*

Please replace the last paragraph on p. 2 through the third full paragraph on p. 4 with the following:

Websites like Genecards (Rebhan, M et al, Bioinformatics 14(8)656-64, 1998) (~~http://nciarray.nci.nih.gov/cards/~~ <http://nciarray.nci.nih.gov/cards/>) provide a database of human genes, their products and their involvement in diseases. However, Genecards only offers information about the functions of all human genes that have an approved symbol, and a few selected others. Again this information can only be accessed one gene at a time, and the annotation cannot be downloaded in any useful format for working with a large gene collection. DRAGON (Bouton CM et al, Bioinformatics 16(11)1038-9, 2000)(~~http://207.123.190.10/dragon.htm~~ <http://207.123.190.10/dragon.htm>) lets the researcher do a keyword search on multiple databases at one time, but the output is a list of accession numbers and definitions in text format, which is not linked to any of its annotations. The tool does not let the researcher select entries from the keyword search. It does not allow moving between pages and merge lists obtained from different keyword searches. As a result DRAGON does not help in systematically compiling a large gene collection. Further, DRAGON does not include important databases like GenBank and LocusLink that are the most commonly used databases for searching candidate genes. None of these tools helps in eliminating sequence redundancies within the lists. Databases like LocusLink and Genecards attempt to integrate the unique characteristics from various databases and provide a broad summary on a single gene

A2



PAGE 2/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

basis. Nevertheless they do not help in annotating a large gene collection. There is a need for a tool that comprehensively gathers annotation related to all these elements in one place. The annotation tool of DRAGON only combines information from UniGene, Swissprot, Pfam and KEGG pathway database with 17 fields of annotation. However these fields do not include important fields like repeat, SNP, pathways, clones, etc. which would be of great value. Additionally including a number (expression data for microarrays, purity of repeats for polymorphism) in the final annotation table would make it convenient for the user to extract information from the table. With more and more gene collections, it is also required to combine several collections of genes, obtained from different sources.

The production of DNA microarrays can be divided into four stages: a. Selection of array elements and design of the probe DNA; b. Preparation of the probe DNA; c. Preparation of a suitable design substrate to spot the probes on; d. Deposition of array elements. The selection of array elements for microarrays involves assembling a large gene collection. It would be very valuable if the same tool (to compile a large gene collection) could be used to further design primers, look for commercially available clones (expression microarrays) and design resequencing probes (resequencing microarrays). Once the genes are spotted on the microarray and hybridized to fluorescent labeled probes, there are a number of software programs that help in conversion of the fluorescence of the scanned image to numbers, using complex mathematical corrections to extract signal from background noise. e.g. Genepix(~~http://www.axon.com/GN_GenePixSoftware.html~~ <http://www.axon.com/GN_GenePixSoftware.html>) and ArrayVision (~~http://imaging.brocku.ca/products/Arrayvision.htm~~ <http://imaging.brocku.ca/products/Arrayvision.htm>). These numbers indicate level of expression. Other programs such as GeneSpring (Silva et al, HMS Beagle: The BioMedNet Magazine Issue 82, 2000), Cluster Treeview (Eisen MB et al, Proc Natl Acad Sci U S A 95) and Spotfire (~~http://www.spotfire.com~~ <http://www.spotfire.com>), help in the analysis by clustering the data together using various methods based on K-means, hierarchal or self-organizing maps. Clustering algorithms use the expression level data to group the various elements on the array. It would also be very useful to view the elements of the array with their complete annotation and overlay the expression level data on top of it. The data could further be selectively viewed by

A2



PAGE 3/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

sorting on various annotation fields and the expression level data. This approach could be useful to view any large gene collection in general. With the increasing number of microarray experiments, it would be valuable to compare elements between different microarrays considering that fragments of the same gene might be represented by different sequence identifiers. For example, two different accession numbers might belong to the same UniGene cluster, representing the same gene. An artifact sometimes observed in the results obtained from an expression profiling microarray experiment is that some sequences might hybridize to other sequences to which they are significantly similar. This leads to false positive results after a microarray experiment. Although Human Cot DNA is often used to prevent non-specific hybridization by blocking simple repetitive elements in genomic DNA, as shown in experiments to study cross-hybridization, Human Cot DNA is not very effective in preventing cross hybridization. ARROGANT computationally estimates the amount of cross hybridization for each sequence and tags potential genes as possible candidates for cross hybridization.

Several computational tools and databases are available which may be used in the development of the code for working with large gene collections. Some of them are discussed here in brief.

A2

1. PRIMO: PRIMO (Li et al, Genomics 40(3) 476-85,1997) is a code that was developed to design primers for large-scale DNA sequencing projects. PRIMO designs primers (short sequences typically 20 bases long), which are used to amplify sequences (0.4 KB- 2 KB) using PCR. PRIMO can be made to design primers to amplify a specific region. PRIMO can be run in batch mode and the region for the design of primers for each sequence can be specified separately. The parameters file (including parameters like oligo length, melting temperatures etc.) can be altered. The code is written in ANSI C and is available locally on a HP/UX computer. The code has been successfully used to design primers for the past couple of years and is available on the web at ~~http://atlas.swmed.edu~~ <http://atlas.swmed.edu>. This makes PRIMO a very important tool to design primers to amplify a large number of sequences simultaneously.

2. BLAST: BLAST(Basic Local Alignment Search Tool) is an alignment tool to search for similar sequences (protein or DNA) developed by NCBI (Altschul et al, Journal of Molecular Biology 215(3)4-3-10,1990). It is available at ~~http://www.ncbi.nlm.nih.gov/BLAST/~~ <http://www.ncbi.nlm.nih.gov/BLAST/>. ARROGANT uses the BLAST output to estimate



PAGE 4/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

A2

cross-hybridization for microarrays. Each element on the array is BLASTed against the entire UniGene database and the BLAST output is parsed to detect 65 contiguous hydrogen bond overlaps, used as a threshold for cross-hybridization.

Please replace the section entitled "Section 1" bridging p. 16 and 17 with the following:

Section 1: Introduction to ARROGANT

ARROGANT is a database driven tool developed to compile, annotate and merge large gene collections. NCBI, KEGG, Research Genetics and other custom databases have been implemented locally since they were the most commonly used databases and were found to extensively cover various items of information related to sequences. The local implementation of various databases and tools (e.g. PRIMO, BLAST) makes ARROGANT independent of other applications and significantly improves its performance. The modular design facilitates addition of new databases with relative ease. ARROGANT has three modes of operation: 1. Design mode (~~http://arrogant.swmed.edu/index1.asp~~ <http://arrogant.swmed.edu/index1.asp>) 2. Analysis mode (~~http://arrogant.swmed.edu/index2.asp~~ <http://arrogant.swmed.edu/index2.asp>) 3. Merge gene collections mode (~~http://arrogant.swmed.edu/index3.asp~~ <http://arrogant.swmed.edu/index3.asp>). The design mode includes keyword searching for compiling gene collections and helps in the design of expression and/or resequencing microarrays. ARROGANT facilitates the design of resequencing and/or expression microarrays by looking for commercially available clones, designing primers and designing probes for resequencing. The analysis mode annotates large gene collections and estimates cross-hybridization for microarrays. When used for microarrays, ARROGANT takes over where ratios or clustering of sequences finishes to provide important data about genes and enables researchers to get a global view. ARROGANT has been used to pre-compute annotation for a large number of gene collections (~~http://arrogant.swmed.edu/precompute.asp~~ <http://arrogant.swmed.edu/precompute.asp>), and the results are stored in the database. This allows quick retrieval of the data and lets the researcher dynamically sort the annotation table. The merging gene collection mode is used to avoid duplicates and redundancies in collections. ARROGANT provides a web based interface and hyperlinks various fields displayed in all the

A3



PAGE 5/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

A3

three modes.

Please replace the first full paragraph on p. 22 with the following:

A4

4. Select Databases: (At least one must be selected) Multiple databases may be selected at one time; options include: a. GenBank; b. UniGene; c. LocusLink; d. KEGG; e. Research Genetics clone database (~~http://www.resgen.com~~ <http://www.resgen.com>). Any combination of the above databases may be used.

Please replace section 3.5.6 bridging p. 23 and 24 with the following:

A5

3.5.6 Design of Primers: ARROGANT uses a code called PRIMO available at ~~http://atlas.swmed.edu~~ <http://atlas.swmed.edu>. The code has been successfully used to design primers for a large number of PCR reactions. PRIMO uses a 'regions' file to design primers and amplify the specified region of interest. ARROGANT creates the 'regions' based on the user's selection to design primers either in the 3' or random region. ARROGANT lets the user modify the parameters used for the design of primers. The parameters include: 1. Oligo length: Length of the primer to be designed which is typically around 20 bases. 2. Tm: Melting temperature to be used for PCR reactions. 3. Number of primers to select (per direction) : Number of forward and reverse primers to select (default = 1). Fig 8 shows a flowchart for primer design.

Please replace the section 4.1.1 on p. 30 through section 4.1.6 on p.32 with the following:

A6

4.1.1 GenBank: GenBank, an annotated collection of all publicly available DNA sequences provided by NIH, is the biggest and the most used publicly available database (Nucleic Acids Research 2000 Jan 1;28(1):15-8). There are approximately 10,897,000 sequence records as of February 2001 (~~http://ncbi.nlm.nih.gov~~ <http://ncbi.nlm.nih.gov>). The complete release notes for the current version of GenBank are available at ~~ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt~~ <ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt>. The GenBank database is the single most important

database to search for possible gene candidates. Each GenBank entry has a unique identifier called accession number. ARROGANT uses accession number as its primary key to link different databases. ARROGANT uses GenBank database in design and analysis mode. GenBank is implemented as a separate database on the server called 'genel fullgenebank' to improve the speed performance, as the database is very large containing approximately 10.8 million entries. The database is implemented as a single table, see Fig 13. ARROGANT GenBank database implemented in SQL Server 7.0 does not include the actual sequence for each entry. This is obtained using the NCBI tools implemented locally on our HP/UX computers. A shell script 'getgb' compares files present locally with its original source on the web and downloads only the ones not existing or having a different file size from ftp://ncbi.nlm.nih.gov/genbank <ftp://ncbi.nlm.nih.gov/genbank>. The files are unzipped, combined into one huge file, split into smaller files of approximately equal sizes and then reformatted and can then be directly imported into the database using the 'bulk insert' script.

4.1.2 UniGene: UniGene partitions GenBank EST sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that presumably represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. The UniGene database was chosen to be a part of ARROGANT (see Fig 14) for the following reasons: 1. Avoid Redundancy: ARROGANT uses UniGene database to avoid redundancies by not including sequences having different accession numbers but representing the same UniGene cluster. ARROGANT uses this in the merging gene collection mode to combine different lists into one unique collection. 2. The UniGene database includes gene sequences as well as hundreds of thousands of expressed sequence tag (EST) sequences. 3. Additional Annotation: Provides additional annotation for a given gene sequence, e.g. cDNA source, which is used to look for keywords (design mode) and annotate gene collection (analysis mode). As a result UniGene database is used in all the three modes by ARROGANT. Perl scripts combine similar files (ftp://ncbi.nlm.nih.gov/repository/UniGene/ <ftp://ncbi.nlm.nih.gov/repository/UniGene/>) of different organisms together, convert the files into various files of specific format which can be imported directly into the database tables using the import function in SQL Server 7.0.



PAGE 7/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

4.1.3 LocusLink: LocusLink is NCBI's attempt to integrate and provide a single query interface to clustered sequences and make available descriptive information about genetic loci. However, LocusLink does not provide annotation to a collection of genes. ARROGANT extends its capabilities by incorporating LocusLink database. Sequence accessions include a subset of GenBank accessions for a locus, as well as a new type, the NCBI Reference Sequence (RefSeq). LocusLink provides a reference sequence for each locus cluster. LocusLink database is used by ARROGANT in the design and analysis mode, see Fig 15. Series of Visual Basic executables import files into the database, downloaded from NCBI (~~ftp://ncbi.nlm.nih.gov/refseq/LocusLink/LL_tmpl~~ <ftp://ncbi.nlm.nih.gov/refseq/LocusLink/LL_tmpl>).

4.1.4 KEGG Genome and Pathway Database: ARROGANT not only combines different databases from NCBI but also uses the KEGG databases. Kyoto Encyclopedia of Genes and Genomes (KEGG) makes available, information pathways consisting of interacting molecules or genes by using the current knowledge of molecular and cellular biology (Kanehisa, M., Oxford University Press 2000). In addition KEGG database also provides additional annotation used by ARROGANT to look for keywords and annotate gene sequences. As a result KEGG database is used by ARROGANT in both design and analysis mode, see Fig 16. The files downloaded from KEGG (~~ftp://kegg.genome.ad.jp/genomes/genes/~~ <ftp://kegg.genome.ad.jp/genomes/genes/>) are combined as one, split into smaller files and the Visual Basic executable is used to update the tables. A file containing additional pathway information is used (~~ftp://kegg.genome.ad.jp/pathways/map_title.tab~~ <ftp://kegg.genome.ad.jp/pathways/map_title.tab>).

4.1.5 HomoloGene: The HomoloGene database provides homologs /orthologs, which is used as a field in the annotation of large gene collection by the analysis mode, see Fig 17. It primarily uses the UniGene cluster identifier to search for homologs / orthologs. Accession numbers and LocusLink identifiers may also be used. HomoloGene uses nucleotide sequence comparisons to calculate orthologs and homologs, between all UniGene clusters by each pair of organisms. The HomoloGene database is downloaded from ~~ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/humig.ftp~~



PAGE 8/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

<ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/hmlg.ftp>. Perl scripts format the downloaded file, which is further imported into the database. Special character '^' is used as the delimiter to import the file into the database.

4.1.6 Research Genetics Clone Database: Research Genetics commercially distributes selected clones from the IMAGE consortium. The catalog of clones available at Research Genetics can be downloaded at ~~ftp://ftp.resgen.com/pub/sv_libraries/RG_Hs_seq_ver_101100.txt~~ <ftp://ftp.resgen.com/pub/sv_libraries/RG_Hs_seq_ver_101100.txt>. The catalog contains annotation related to the clones like accession number, gene name, cluster ID, insert size, markers, etc. ARROGANT stores this catalog locally in the database, which is used to find commercially available clones and search for candidate genes in the design mode, see Fig 18.

A6

---

Please replace section 6.3 on p. 41 through the first partial paragraph on p. 45 with the following:

---

6.3 ARROGANT used for identifying and annotating genes for polymorphism discovery to link to cardiac diseases for PGA: The Program for Genomic Application (PGA) is a nationwide attempt to use genomic and proteomic methods to study and investigate cellular responses to injury and inflammation. The program endeavors to identify the genes and proteins involved in these responses. ARROGANT was used to both recommend new candidate genes for PGA as well as annotate the current PGA list of 253 genes. The ability of ARROGANT to find potential candidates was tested by comparing the list obtained using keyword search with the current list of genes. The list of keywords compiled by researchers participating in PGA was as follows:

A7



PAGE 9/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

| | |
|---|---|
| hyperlipidemia | arteriosclerosis |
| low density lipoproteins | cholesterol |
| dietary responsiveness | inflammation |
| high density lipoproteins | cytokine |
| coronary calcification | orphan receptor |
| insulin resistance | cardiac failure |
| cardiac hypertrophy | signal transduction |
| coronary artery disease | G-protein |
| coronary atherosclerosis | |

ARROGANT found 3,789 genes associated with the above keywords. There were 13 genes found in common with the current PGA list of 253 genes. This demonstrated the keyword search capability of ARROGANT to look for potential candidates associated with keywords. The newly compiled list was annotated using the analysis mode and is available on the web at: ~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=40710~~ <http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=40710>. ARROGANT was also used to annotate the current PGA list of 253 genes.

The ability of ARROGANT in the analysis mode to accept a list of genes tab delimited with a number was used to assign priority levels to the genes: 2- Highest priority, 1- Moderate priority and 0- Low priority. The annotated table is available on the web at: ~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=44082~~ <http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=44082>.

A7

6.4 ARROGANT used in the study of Robert's Syndrome: Robert's Syndrome is a genetic disorder caused by chromosome damage during cell division, and characterized by loss of limb bones, cleft palate, heart defects and abnormalities of the abdominal organs. ARROGANT was used to find new potential candidate genes for Robert's syndrome using keywords:



PAGE 10/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

| | |
|---|---|
| Robert syndrome | hypoplastic nasal and auricular cartilage |
| Roberts syndrome | atrial septal defect |
| Robert's syndrome | patent ductus arteriosus |
| Pseudothalidomide syndrome | polycystic kidneys |
| SC phocomelia syndrome | fused kidneys |
| heterochromatin | horseshoe kidneys |
| Heterochromatic repulsion | micronucleation |
| Heterochromatic splaying | enlargement of the phallus |
| Premature centromere separation | absent nails |
| premature separation | ICF syndrome |
| Tetraphocomelia | Centromeric instability immunodeficiency |
| Limb reduction | syndrome |
| hypoplastic | MECP2 |
| Long bone | Methyl binding protein |
| Aneuploidy | Hypomethylation |
| Craniofacial | Hypermethylation |
| Oxycephalic | Demethylation |
| aplasia of the fibula | demethyltransferase |
| bilateral clubfoot | Methylation |
| absence of radii | methylase |
| cleft lip and palate | mSIN3A |
| oligodactyly | Histone |
| microcephaly | Histone acetylation |
| exophthalmus | Histone acetylase |
| hypertelorism | Histone deacetylase |
| corneal clouding | TAR syndrome |
| hemangiomas | |

A7

ARROGANT found 6,326 genes, which were further annotated using the analysis mode. The



PAGE 11/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

results are available on the web at:
~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=12345~~
<http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=12345>. A separate list of 16 gene names found to be important in the study of Robert Syndrome was obtained. The accession numbers for these 16 genes were determined using ARROGANT. It was found that there was one gene in common between the two lists. This again demonstrated the utility of ARROGANT to look for and identify candidate genes associated with keywords. The list of 16 genes was also annotated using the analysis mode and the results are available on the web at
~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=39613~~
<http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=39613>.

6.5 ARROGANT used to annotate genes on commercial DNA chips: ARROGANT was used in the analysis mode to annotate various microarrays available from Affymetrix (Santa Clara, CA) to help the researcher view the results obtained from the expression studies in a convenient manner. This provides the researcher a group of genes having particular characteristics together and helps in making important observations. The following commercial (Affymetrix) human and mouse microarrays were analyzed.

1. Human HUG95 microarray: This microarray consists of 12,454 different elements. The annotated list is available on the web at
~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=11111~~
<http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=11111>.

2. Rat RG-U34 microarray: This consists of 1,322 genes from Rat genome. The results are available on the web at
~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=57860~~
<http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=57860>.

6.6 ARROGANT used to annotate genes on chromosome 3p: ARROGANT was used to identify genes commonly mutated or whose expression is deregulated in human lung and breast cancers. Although several regions of loss occur on multiple chromosomes it was observed that allele loss in the chromosome 3p21.3 area was the earliest pre-malignant change so far detected in lung



PAGE 12/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12

cancer development (~~http://www.utsouthwestern.edu/cancer/Research/3p21_intro.htm~~ <http://www.utsouthwestern.edu/cancer/Research/3p21_intro.htm>). ARROGANT was used to annotate the 32 genes on chromosome 3p thought to be important in causing lung cancer. The results are available at: ~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=40357~~ <http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=40357>.

6.7 ARROGANT used to analyze human microarrays: Our laboratory has developed a human cDNA microarray, which consists of 10,000 clones from Research Genetics. Many laboratories in UTSW (University of Texas Southwestern Medical Center at Dallas) are using this microarray for various research studies like cancer, aging, etc. ARROGANT provides annotation for all the genes as one table. The researchers can overlay their expression level data on this table, which would help them make important observations. For example, the researcher could look at the pathways for all the highly expressed genes and also know their position in the genome. Further the researcher could also sort the data using ARROGANT to bring the interesting genes on top of the table. ARROGANT annotation of the human 10,000 array is available on the web at ~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=60110~~ <http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=60110>. ARROGANT also annotated our earlier human array consisting of 4,200 elements and the results are available at ~~http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=60718~~ <http://ARROGANT.swmed.edu/myweb/hideandsort.asp?txt_array=60718>.

AM



PAGE 13/16 * RCVD AT 12/9/2003 2:47:00 PM [Eastern Standard Time] * SVR:USPTO-EFXRF-2/4 * DNIS:7467239 * CSID:6503434342 * DURATION (mm-ss):08-12